Combination of classifiers with optimal weight based on evidential reasoning

Zhun-Ga Liu, Quan Pan, Jean Dezert, Arnaud Martin

Combination of classifiers with optimal weight based on evidential reasoning

Zhun-ga Liu[1], Quan Pan[1], Jean Dezert[2], Arnaud Martin[3]
1. School of Automation, Northwestern Polytechnical University, Xi'an, China.
Email: liuzhunga@nwpu.edu.cn
2. ONERA - The French Aerospace Lab, F-91761 Palaiseau, France.
Email: jean.dezert@onera.fr
3.IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France.
Email: Arnaud.Martin@univ-rennes1.fr

***Abstract*—In pattern classification problem, different classifiers learnt using different training data can provide more or less complementary knowledge, and the combination of classifiers is expected to improve the classification accuracy. Evidential reasoning (ER) provides an efficient framework to represent and combine the imprecise and uncertain information. In this work, we want to focus on the weighted combination of classifiers based on ER. Because each classifier may have different performance on the given data set, the classifiers to combine are considered with different weights. A new weighted classifier combination method is proposed based on ER to enhance the classification accuracy. The optimal weighting factors of classifiers are obtained by minimizing the distances between fusion results obtained by Dempster's rule and the target output in training data space to fully take advantage of the complementarity of the classifiers. A confusion matrix is additionally introduced to characterize the probability of the object belonging to one class but classified to another class by the fusion result. This matrix is also optimized using training data jointly with classifier weight, and it is used to modify the fusion result to make it as close as possible to truth. Moreover, the training patterns are considered with different weights for the parameter optimization in classifier fusion, and the patterns hard to classify are committed with bigger weight than the ones easy to deal with. The pattern weight and the other parameters (*i.e.* classifier weight and confusion matrix) are iteratively optimized for obtaining the highest classification accuracy. A cautious decision making strategy is introduced to reduce the errors, and the pattern hard to classify will be cautiously committed to a set of classes, because the partial imprecision of decision is considered better than error in certain case. The effectiveness of the proposed method is demonstrated with various real data sets from UCI repository, and its performances are compared with those of other classical methods.**

Keywords: evidential reasoning, Dempster-Shafer theory (DST), combination rule, classifier fusion, belief functions.

I. Introduction

Ensemble classifier has been considered as an efficient way to achieve the highest possible accuracy in pattern classification problem [1]–[4]. The different classifiers usually provide some complementary knowledge about the query pattern, and the ensemble system can take advantage of such complementarity to improve the classification accuracy. Thus, the ensemble classifier produces a higher accuracy rate than the best classifier in general. The ensemble classifier broadly consists of two parts 1) the choice and implementation of classifiers, like boosting and bagging, and 2) the combination of classifiers in a particular way. In this work, we want to focus on the second part about how to efficiently combine the classifiers in the multi-class problem. The classifier fusion methods can be generally divided into three groups according to the type of the individual output [3], *i.e.* crisp labels, class rankings and soft outputs. The class labels are often combined by the voting methods. The class set reduction/reordering methods are usually applied to merge the class rankings. The soft output (*e.g.* probability, fuzzy membership, belief functions) providing more useful classification information can be combined by Bayesian rule [5], fuzzy integrals [6] and ER [7], [8].

In the complex pattern classification problem, the classification result produced by single classifier may be quite uncertain due to the limitation of observed attributes. Evidential reasoning also called Dempster-Shafer theory (DST) or belief function theory [9]–[12] provides a theoretical framework to model and combine the uncertain information [14]. In the combination of sources of evidence (*i.e.* classifiers), the reliability of each source can be considered via Shafer's discounting operation in ER. Particularly, the contextual discounting operation (*i.e.* an extension of Shafer's discounting operation) has been further developed by Mercier in [27], and it allows to take into account more refined reliability knowledge conditionally on different hypotheses regarding the variable of interest. ER has already been used successfully in many fields of applications, *e.g.* information fusion [15], pattern recognition [16]–[21], parameter estimation [23]–[26], etc. Some evidential classification methods, *e.g.* Evidential K-nearest Neighbors (EKNN) [22], Evidential Neural Network (ENN) [16], have been proposed by Denœux based on DST, and these evidential methods can well handle the uncertainty in pattern classification for achieving a good performance. We have developed several credal classifiers to further characterize the partial imprecise information in different cases [17], [18], and our previous methods allow the object to belong to not only singleton classes but also meta-classes (*i.e.* the disjunction of several singleton classes) with different masses of beliefs.

ER has been used for classifier fusion to improve the ac-

curacy. Three classifier fusion techniques including Sugeno's fuzzy integral, the possibility theory and DST are applied for Automatic Target Recognition (ATR) to improve the accuracy of individual classifiers in [30], and it shows that DST usually achieves the best performance. In [31], an interesting Basic Belief Assignment (BBA) generation method is presented for the combination of multiple classifiers based on DST. The class decision of each classifier is described by a simple BBA, and the mass of belief focusing on the singleton class is calculated according to the distance between the classifier output and the reference vector, which is obtained by minimizing the mean square errors between combined classifier outputs and the target values. A class-indifferent method is developed in [8] for the classifier fusion based on DST, and each classifier output is represented by evidential structures of triplet and quadruplet, which can distinguish the important classes from the trivial ones. The ignorant elements have been employed to model the unknown and uncertain class decisions. The parameterized t-norm based combination rules are introduced in [7] for the fusion of non-independent classifiers under belief functions framework, and it behaves ranging between Dempster's rule and the cautious rule by tuning the parameters, which are optimized by minimizing an error criteria. There are two fusion strategies (*i.e.* a single combination rule and a two-step fusion method) investigated for obtaining the optimal combination scheme. In [28], postal address recognition method is developed based on the fusion of the outputs from multiple Postal Address Readers (PAR, regarded as classifier) using transferable belief model (TBM) [12], and the PAR outputs can be properly converted into belief functions according to the confusion matrix reflecting the classification performance of PAR.

In the fusion of multiple classifiers, each classifier may play a different role, since they often have different classification performances. Thus the classification accuracy could be further improved by assigning the appropriate weights to classifiers in the fusion. ER provides an efficient tool for handling the uncertainty in the multiple sources of information fusion, and evidence discounting operation can well control the influence of each source (*i.e.* classifier) in the fusion according to the given weights. Hence, we want to develop a new weighted combination method for different classifiers based on ER to enhance the classification accuracy.

The weighted averaging combination rule has been widely applied in classifier fusion, and the classifier weight is often determined depending on the individual accuracy [32]. The fusion method can improve the accuracy with respect to the individuals mainly because of the complementarity of classifiers. Nevertheless, the important complementary knowledge cannot be efficiently taken into account if the classifier weight is calculated only by the accuracy. There also exist some other methods for optimizing the weights of classifiers, but these methods are not applicable for ER combination scheme. Moreover, the training patterns are often considered equal in the calculation of classifier weight[1]. In fact, the tuning of classifier weight has in general a very little influence on the class decision making for the pattern that can be easily classified. Whereas, the class decision for the pattern hard to classify is usually sensitive to the changes of classifier weight. As a result, the training patterns cannot be equally treated in the optimization of classifier weight. In the class decision making step, the hard classification usually assigns the object to a singleton class with the biggest probability value, but this strategy may cause high risk of error especially for the object with high uncertainty of classification. Hence, it seems interesting to develop a cautious decision making strategy to reduce the number of classification errors.

We propose a new weighted combination method for multiple classifiers working with different features (*i.e.* attributes) of pattern. The weight of each classifier is optimized by minimizing an error criteria. A confusion matrix, which characterizes the probability of the object belonging to one class but classified to another class, is introduced to further improve the classification performance. Moreover, the training patterns are given different weights in the parameter optimization based on the distances of their classification results to the truth. Thus, the weights of training patterns and the fusion parameters (*i.e.* the weights of classifiers and confusion matrix) are iteratively optimized for achieving the best possible result.

This paper is organized as follows. After the brief introduction of background knowledge of ER in section II, the combination of classifiers with optimal weights is presented in detail in section III. Then the cautious decision making strategy is given in section IV for the final classification. The performance of proposed method is tested in section V and compared with other related fusion methods before giving our concluding remarks in section VI.

II. Background knowledge of evidential reasoning

Evidential reasoning (ER) [9]–[12] also called belief function theory or Dempster-Shafer theory (DST) works with a frame of discernment as $\Omega = \{\omega_1, \ldots, \omega_c\}$ consisting of c exclusive and exhaustive hypotheses (*i.e.* classes) $\omega_i, i = 1, \ldots, c$. The basic belief assignment (BBA) in ER is defined over the power-set of Ω denoted by 2^Ω, which is composed of all the subsets of Ω. The power-set 2^Ω contains $2^{|\Omega|}$ elements including the empty set as $2^\Omega = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_c\}, \{\omega_1, \omega_2\}, \ldots, \Omega\}$. The cardinality of a set as $|A|$ denotes the number of elements included in A.

A BBA is represented by a mass function $m(.)$ from 2^Ω to $[0, 1]$ such that $m(\emptyset) = 0$ and $\sum_{A \in 2^\Omega} m(A) = 1$. All the elements $A \in 2^\Omega$ such that $m(A) > 0$ are called the focal elements of the BBA $m(.)$. The set $\mathbf{K}(m) \triangleq \{A \in 2^\Omega \mid m(A) > 0\}$ of all focal elements of the BBA $m(.)$ is called the core of $m(.)$. The object is allowed to belong to not only singleton classes (*e.g.* ω_i), but also any subsets of Ω (*e.g.* $A =$

[1]In the Boosting approach, the training patterns are assigned with different weights, but this method works with quite distinct principle. The different classifiers are closely relevant in Boosting, whereas the classifiers are considered independent in this work.

$\{\omega_i, \omega_j\}$ with different masses of belief. The total ignorance is represented by Ω.

The lower and upper bounds of probability associated with a BBA respectively correspond to the belief function $Bel(.)$ and the plausibility function $Pl(.)$ [10] defined by $\forall A \subseteq \Omega$

$$Bel(A) = \sum_{B \in 2^\Omega | B \subseteq A} m(B). \tag{1}$$

$$Pl(A) = \sum_{B \in 2^\Omega | A \cap B \neq \emptyset} m(B). \tag{2}$$

In pattern classification problem, the soft output of each classifier can be considered as one source of evidence represented by a BBA, and the probabilistic output is considered as the simple Bayesian BBA. Dempster's rule (also called DS rule) remains very popular in the combination of multiple sources of evidence, because it is commutative and associative, which makes it very appealing from implementation standpoint. Let us consider two BBA's $\mathbf{m}_1$ and $\mathbf{m}_2$ ($\mathbf{m}_i \triangleq m_i(.)$ for conciseness) defined over 2^Ω. The combination of $\mathbf{m}_1$ and $\mathbf{m}_2$ by DS rule is defined by $B, C \in 2^\Omega$

$$m(A) = \mathbf{m}_1 \oplus \mathbf{m}_2(A) = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \mathcal{K}}, \forall A \in 2^\Omega \setminus \{\emptyset\}. \\ 0, \quad \text{if} \quad A = \emptyset. \end{cases} \tag{3}$$

where $\mathcal{K} = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ measures the degree of conflict between the BBA's. The denominator $1 - \mathcal{K}$ is used for the normalization of combination result. It is worth noting that DS rule is applicable only if $\mathcal{K} = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \neq 1$. The vacuous BBA $m(\Omega) = 1$ plays a neutral role in DS fusion, and any BBA remains unchanged when combined by DS rule with the vacuous BBA.

In the high conflicting cases and some special cases [33], [34], DS rule may produce unreasonable results due to the redistribution way of conflicting masses $\mathcal{K}$. Thus, a number of alternative combination rules have emerged to overcome the limitations of DS rule, such as Yager's rule, Dubois-Prade (DP) rule, and Proportional Conflict Redistribution (PCR6) rule [34]. Unfortunately these rules are not associative and that is why they are not so appealing from the implementation standpoint in the real applications. DS rule will be used in this work to combine the classification results provided by different classifiers because its associativity property makes it easier to implement than other fusion rules.

The classifiers to combine may have different reliabilities because they usually have different abilities of classification. A particular discounting operation has been introduced by Shafer in [10] for the combination of sources of information with different reliabilities, and it discounts the masses of all focal elements by a discounting (weighting) factor $\alpha \in [0, 1]$ to the total ignorance. By doing this, one can efficiently control the influence of each classifier in the fusion. More precisely, the discounted mass is obtained by the formula

$$\begin{cases} {}^\alpha m(A) = \alpha \cdot m(A), A \subset \Omega, A \neq \Omega. \\ {}^\alpha m(\Omega) = 1 - \alpha + \alpha \cdot m(\Omega). \end{cases} \tag{4}$$

If the source of evidence is considered completely reliable, one takes $\alpha = 1$. Then, the BBA remains the same after the discounting as ${}^\alpha m(.) = m(.)$. If the evidence is not reliable at all, we set $\alpha = 0$, and the mass values of all the focal elements will be discounted to the ignorance as ${}^\alpha m(\Omega) = 1$.

In certain cases, the reliability of each source of evidence can be expected to depend on the truth of the variable of interest. In order to take into account such refined reliability knowledge (*i.e.* conditionally on values taken by the variable of interest), the contextual discounting operation has been proposed by Mercier in [27], and the learning of discount rates is also addressed by minimizing the discrepancy between plausibility and observations. This contextual discounting operation can properly redistribute the masses of belief according to the reliability vector. The set of tools has been enlarged in [29] to deal with the contextual knowledge about the source quality in terms of relevance and truthfulness based on belief function theory. The practical means to learn the contextual knowledge from available labeled data are also introduced in [29], and it makes the correction mechanisms interesting and useful in practice.

In this work, we consider the simple case that each classifier is given only one weight as normally done in the classifier fusion problem, and Shafer's discounting operation will be adopted. The combination of a pair of discounted BBA's (*i.e.* ${}^{\alpha_1}\mathbf{m}_1$ and ${}^{\alpha_2}\mathbf{m}_2$) by DS rule with discounting factors α_1 and α_2 is directly given for the convenience by ${}^\alpha\mathbf{m} = {}^{\alpha_1}\mathbf{m}_1 \oplus {}^{\alpha_2}\mathbf{m}_2$. For $B, C \in 2^\Omega$,

$$\begin{cases} {}^\alpha m(A) = \frac{\sum_{B \cap C = A} \alpha_1 \alpha_2 m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} \alpha_1 \alpha_2 m_1(B) m_2(C)}, \forall A \in 2^\Omega \setminus \{\emptyset, \Omega\}. \\ {}^\alpha m(\Omega) = \frac{\delta}{1 - \sum_{B \cap C = \emptyset} \alpha_1 \alpha_2 m_1(B) m_2(C)}. \\ {}^\alpha m(\emptyset) = 0. \end{cases} \tag{5}$$

where $\delta = 1 - \alpha_1[1 - m_1(\Omega)] - \alpha_2[1 - m_2(\Omega)] + \alpha_1\alpha_2[1 - m_1(\Omega) - m_2(\Omega) + m_1(\Omega)m_2(\Omega)]$.

III. Optimal combination of multiple classifiers

Let us consider one object (say $\mathbf{y}$) being classified over the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$ according to the proper combination of n classifiers (*i.e.* $\mathcal{C}_1, \ldots, \mathcal{C}_n$), which are respectively trained by a set of labeled patterns (*i.e.* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$) on n different attribute (feature) spaces as $\mathbb{S}_1, \ldots, \mathbb{S}_n$. The class label of pattern $\mathbf{x}_k$ is represented by $L(\mathbf{x}_k)$. It is assumed that each classifier $\mathcal{C}_l, l = 1, \ldots, n$ produces soft output represented by the probabilistic membership (for classifier under probabilistic framework) or belief degree (for evidential classifier working with belief functions) of the object belonging to each class. The probabilistic output can be always interpreted as Bayesian BBA [10], and the output of evidential classifier is also denoted by BBA consisting of some singletons $\omega_i \in \Omega, i = 1, \ldots, c$ and the total ignorance Ω as focal elements. Evidential reasoning (ER) providing an efficient tool to deal with the uncertain information is employed for combining classifiers.

A. Combination of classifiers with different weights

The classifiers to combine are learnt based on different attribute knowledge, and they may own different abilities of classification. Thus, each classifier will be given an appropriate weight in the fusion in order to achieve the best possible classification result. In the traditional methods, the classifier weight is usually determined according to its performance (*e.g.* accuracy) on the training data set. The lower accuracy, the smaller weight. By doing this, it can reduce the influence of the classifier with low accuracy. In such methods, the weight of individual classifier is calculated separately regardless the complementarity among the different classifiers. Nevertheless, the proper combination of different weak but complementary classifiers (with low accuracy) may still produce good results if we can take fully advantage of their complementary knowledge via the fusion procedure.

In this work, the classifier weight will be calculated based on the optimization procedure with DS combination, and the optimal weight should make the combination results as close as possible to the truth for the training patterns. This is a classical optimization strategy in classification problems [16], [27], [31]. Hence, the optimal classifier weighting vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ ($\alpha_l \in [0,1], l = 1, \ldots, n.$) can be estimated by minimizing the distance between the combination result and the true class of the training patterns. Jousselme's distance $d_J(\cdot,\cdot)$ [13] taking into account both the differences of mass values and the intersection of focal elements is often used to measure the distance of a pair of BBA's, and it is employed here. Therefore, one must calculate

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{k=1}^{K} d_J(\bigoplus_{l=1}^{n} {}^{\alpha_l}\mathbf{m}_{kl}, \mathbf{T}_k). \tag{6}$$

The output of classifier $\mathcal{C}_l$ with respect to pattern $\mathbf{x}_k$ is represented by the BBA as $\mathbf{m}_{kl}$. The truth of classification of the training pattern $\mathbf{x}_k$ with label $L(\mathbf{x}_k)$ is characterized by the binary vector[2] $\mathbf{T}_k = [T_{k1}, T_{k2}, \ldots, T_{kc}]$. All components of $\mathbf{T}_k$ are equal to zero but $T_{kt} = 1$ for the class $\omega_t = L(\mathbf{x}_k)$. Jousselme's distance for a pair of BBA's $\mathbf{m}_1$ and $\mathbf{m}_2$ is defined by:

$$d_J(\mathbf{m}_1, \mathbf{m}_2) \triangleq \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)'\mathbf{D}(\mathbf{m}_1 - \mathbf{m}_2)}. \tag{7}$$

where $\mathbf{D}$ is a $2^{|\Omega|} \times 2^{|\Omega|}$ positive matrix. Its components are defined by Jaccard's factors $D_{ij} \triangleq \frac{|A_i \cap B_j|}{|A_i \cup B_j|}$, $A_i, B_j \in 2^{\Omega}$. Because in this work the cores of ${}^{\alpha_l}\mathbf{m}_{kl}$ are restricted only to singletons and to Ω, $\mathbf{D}$ matrix is restricted to a $(|\Omega|+1) \times (|\Omega|+1)$ matrix. This distance measure has been aplied for the decision making in pattern classification problem [41].

The following lemma justifies the use of classifier weight with discounting technique to reduce the errors in classifier fusion.

Lemma 1. Let us consider a frame of discernment $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ and $n \geq 2$ discounted BBA's ${}^{\alpha_l}\mathbf{m}_l$ with cores $\mathbf{K}({}^{\alpha_l}\mathbf{m}_l) = \{\omega_i \in \Omega, \Omega\}$ and with weighting factors (*i.e.* weights of classifiers) $\alpha_l \in [0,1]$ for $l = 1, \ldots, n$. The pattern will be classified based on the resulting BBA as ${}^{\alpha}\mathbf{m}$ obtained by the combination of these n discounted BBA's using DS rule. It is assumed that one pattern to classify truly belongs to ω_t. It is possible to choose the weighting factors $\{\alpha_l, l = 1, \ldots, n\}$ such that $\omega_t = \arg\max_{\omega_i}[{}^{\alpha}m(\omega_i)]$ if $\nexists\ \omega_i \neq \omega_t, m_l(\omega_i) \geq m_l(\omega_t)$ for all $l = 1, \ldots, n$.

Under the above condition, this lemma states that the suitable weighting factors corresponding to the weights of classifiers can be chosen in such a way that the true class (*i.e.* ω_t) of the pattern gets the maximum mass value by the combination of the discounted BBA's with DS rule. Therefore, the proper tuning of classifier weight is an interesting mean to reduce the classification errors in the fusion[3].

The combination of the discounted BBA's with proper weighting factors can produce the correct classification, and the corresponding combination result is generally closer to the truth of classification than the combination result leading to the error with improper weighting factors. The optimal weighting factors will be determined by minimizing the distance between the combination result and the truth using training data, and it will be presented in the sequel.

B. Confusion matrix for belief transformation

The weighting vector $\boldsymbol{\alpha}$ is used to discount the classification results produced by different classifiers in order to control the influence of each classifier in the fusion procedure. Because DS rule is based on conjunctive rule, the product of the mass of non contradicting focal elements is committed to their intersection. For any class, if its plausibility value (upper bound of probability) is nonzero in each BBA, the positive plausibility will still be committed to this class whatever the classifier weight $\boldsymbol{\alpha}$ is. Thus, the positive probabilities (or beliefs) are often committed to multiple classes in the combination result of different classifiers for the uncertain pattern, which truly belongs to only one class, and there usually exists more or less bias between the weighted combination result and the truth. If we want to make the classifier fusion result as close as possible to truth, it is necessary to transfer (redistribute) the beliefs among different classes judiciously. In fact, the use of the classifier weight $\boldsymbol{\alpha}$ only is insufficient for making this judicious belief redistribution. That is why we also need to introduce the confusion matrix, which is justified in the following lemma.

Lemma 2. Let us consider that one pattern $\mathbf{x}_k$ truly belonging to ω_t is classified by combining n pieces of classifier outputs as $\mathbf{m}_l, l = 1, \ldots, n$ with cores $\mathbf{K}({}^{\alpha_l}\mathbf{m}_l) = \{\omega_i \in \Omega, \Omega\}$ and with weighting factors $\alpha_l \in [0,1]$ for $l = 1, \ldots, n$. The combination result of these n discounted BBA's by DS rule is denoted by ${}^{\alpha}\mathbf{m}$. For any values of α_l, $l = 1, \ldots, n$, the inequality $d_J({}^{\alpha}\mathbf{m}, \mathbf{T}_k) > 0$ holds if $\exists\ \omega_g \neq \omega_t \in \Omega$ such that $\prod_{l=1}^{n} Pl_l(\omega_g) > 0$.

Lemma 2 states that if the classifiers to combine commit even little plausibility to a common element ω_g (rather than

[2]For the evidential classifier, the ignorant element Ω is a focal element of the classifier output, and therefore one must include also an extra component $T_{k,c+1} = 0$.

[3]The proof of Lemma 1 can be found in the supplementary materials online.

the true class ω_t) or to the total ignorance Ω, then their combination result by DS rule will never achieve the truth of classification whatever the values of the weighting factors are[4].

In order to make the combination result as close as possible to the truth, we propose to introduce a confusion matrix describing the prior probability of the object belonging to one class when it is classified to another class based on the classifier fusion result. One can transform (correct) the beliefs of the different classes using this confusion matrix to improve the accuracy of the classification. The confusion matrix is denoted $\boldsymbol{\beta} = [\beta_{ij}]_{c\times c}$ (c being the number of classes in the frame of discernment), and each element β_{ij} represents the conditional probability of the object (say $\mathbf{x}$) belonging to class ω_j if it is classified to ω_i according to the combination result of classifiers. More precisely, $\beta_{ij} \triangleq p(L(\mathbf{x}) = \omega_j | \hat{L}(\mathbf{x}) = \omega_i)$, where $\hat{L}(\mathbf{x})$ denotes the estimated class label of the object $\mathbf{x}$ based on the combination of classifiers, and $L(\mathbf{x})$ represents the true label. Of course, the following equality must hold: $\sum_{j=1}^{c} \beta_{ij} = 1$. The weighted combination result of classifiers as ${}^{\alpha}m(.)$ are adjusted using $\boldsymbol{\beta}$ by

$$m(\omega_j) = \sum_{i=1}^{c} {}^{\alpha}m(\omega_i)\beta_{ij}. \tag{8}$$

Hence, the confusion matrix[5] $\boldsymbol{\beta}$ is also included in the objective function which is now expressed by eq. (9).

$$\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\} = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{k=1}^{K} d_J((\bigoplus_{l=1}^{n} {}^{\alpha_l}\mathbf{m}_{kl})\boldsymbol{\beta}, \mathbf{T}_k). \tag{9}$$

subject to constraints

$$\begin{cases} \alpha_l \in [0,1], \quad l = 1,\ldots,n. \\ \sum_{j=1}^{c} \beta_{ij} = 1, \quad i = 1,\ldots,c. \end{cases} \tag{10}$$

Optimal parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be found by minimizing this objective function.

Lemma 3 is given to justify the use of the confusion matrix for improving classification accuracy.

Lemma 3. A set of patterns is classified according to the combination result of classifiers denoted by ${}^{\alpha}m(.)$. Let us consider the patterns belonging to a disjunction of two classes[6] say ω_i and ω_j. A proper confusion matrix $\boldsymbol{\beta}$ can be found under a certain condition to improve the accuracy of the fusion of classifiers.

In Lemma 3, we discuss the conditions of existence of the proper confusion matrix for the correction of BBA's in different cases to show the potential of this correcting step for the further improvement of classification accuracy[7]. The optimal confusion matrix jointly with the classifier weight will be calculated by minimizing the classification result and the ground truth using training data, and it will be explained in the next subsection.

C. Taking into account the pattern weight

In real applications, we usually classify the pattern to the class with the biggest probability or mass of belief. For the pattern easy to classify (*e.g.* each classifier assigns the high probability to the correct class), its classification result will be not very sensitive to the tuning of parameters (*i.e.* classifier weight, confusion matrix) for making the correct classification. Nevertheless, some other patterns with quite uncertain classification results can be hard to classify. Their classification results are usually very sensitive to the tuning of parameters in the fusion, and a small change can turn the correct classification to an error. We must pay more attention to such pattern in parameter estimation for classifier fusion. These uncertain patterns should be assigned with the bigger weights in the parameter optimization procedure than the patterns easy to classify.

The objective function taking into account the pattern weight $\mathbf{w} = (w_1,\ldots,w_K)$ is given by:

$$f = \sum_{k=1}^{K} w_k d_J((\bigoplus_{l=1}^{n} {}^{\alpha_l}\mathbf{m}_{kl})\boldsymbol{\beta}, \mathbf{T}_k). \tag{11}$$

subject to the constraints

$$\begin{cases} \alpha_l \in [0,1], \quad l = 1,\ldots,n; \\ \sum_{j=1}^{c} \beta_{ij} = 1, \quad j = 1,\ldots,c; \\ w_k \in [0,1], \quad k = 1,\ldots,K. \end{cases} \tag{12}$$

Because it is hard to determine the optimal parameters *i.e.* $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)$, $\boldsymbol{\beta} = [\beta_{ij}]_{c\times c}$ and $\mathbf{w} = (w_1,\ldots,w_K)$ by minimizing directly the objective function (11) under the constraints (12), we use an iterative optimization procedure. The detailed calculation of these parameters is presented as follows. At the beginning, each training pattern will be considered with equal weight as $w_k = 1$, for $k = 1,\ldots,K$. Then the classifier weight $\boldsymbol{\alpha}$ and the confusion matrix $\boldsymbol{\beta}$ can be obtained by minimizing the objective function (11). We use the active-set algorithm [36] to solve this optimization problem. One can compute the combination result of the n classifiers for each pattern with the optimized parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then the training pattern weight w_k is modified according to the distance between the combination result and the truth of classification in training data space.

If the combination result is quite close to the ground truth for the labeled training pattern, it implies that the parameters obtained in last step can be tuned in some degree keeping the correct decision for this pattern. Hence, this pattern will receive a small weight. If the distance between the combination result and the ground truth is big, we must assign a bigger weight to this pattern in the optimization procedure.

[4]The proof of Lemma 2 is given in the supplementary materials online.

[5]The masses of beliefs of singleton elements will be redistributed according to the matrix $\boldsymbol{\beta}$. The mass of ignorance in evidential classifier is usually very small, and it will be redistributed in the decision making step as done in transferable belief model (TBM) model [12].

[6]In the classification of uncertain data, the different classes can partially overlap. Each overlapping zone usually contains a few (e.g. two) classes. For simplicity, we just consider here the case of misclassification between two classes ω_i and ω_j. Other misclassified classes can be similarly handled by the corresponding elements in the confusion matrix.

[7]The proof of Lemma 3 can be seen in the supplementary materials online.

Generally, the bigger distance value, the bigger pattern weight. Hence, the pattern weight $w_k \in [0,1]$ should be a monotone increasing function of the distance measure where $d_k \in [0,1]$ as $w_k = f(d_k)$. If the distance d_k is approximately zero, it means the combination result is almost equal to ground truth, and the weight w_k value can be also close to zero. If the distance reaches its maximum value $d_k = 1$, this pattern weight will be considered with the biggest value $w_k = 1$. The slope of increasing for the function $f(.)$ mainly depends on the actual application. Moreover, this function should be simple for the convenience of application. According to this basic principle, the pattern weight can be defined by

$$w_k = d_k^{\lambda}. \tag{13}$$

where $d_k \in [0,1]$ is the Jousselme's distance [40] between the combination result and the target value (*i.e.* truth of classification) of training data involved in (11), and where $\lambda > 0$ is a penalized coefficient which controls the slope of increasing of pattern weight with the increasing of the distance value. The bigger λ, the bigger slope. It will be tuned according to the current context for the global improvement of the classification accuracy.

Once the pattern weight is updated, we will recalculate the classifier weight $\boldsymbol{\alpha}$ and the confusion matrix $\boldsymbol{\beta}$ with this updated pattern weight. Then the corresponding accuracy will be computed according to the combination result. If the accuracy becomes higher than before, this updated $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be adopted. Otherwise, we will still keep the previous estimation of the parameters. Such iterative procedure will be stopped as soon as the accuracy cannot be improved.

The pseudo-code of the new method is given in Table I.

Table I
COMBINATION OF MULTIPLE CLASSIFIERS WITH OPTIMAL WEIGHT

Input: training patterns $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_K\}$ trained classifiers $\mathcal{C}_1, \ldots, \mathcal{C}_n$
Initialization: $\mathbf{w}_0 = ones(1,K), \boldsymbol{\alpha}_0 = ones(1,n)$, $\boldsymbol{\beta}_0 = eye(c), AC_0 = 0, Sgn = 1.$

Implementation: t $\leftarrow$ 0
While {Sgn}
t $\leftarrow$ t+1
Compute $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ to minimize eq. (11) with $\mathbf{w}_{t-1}$;
Compute the classifier fusion result with $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$;
Compute the classification accuracy AC_t for X;
If $AC_t - AC_{t-1} > 0$
Compute pattern weight $\mathbf{w}_t$ using eq. (13);
else
Sgn=0;
EndIf
Endwhile

Output: Classifier fusion result with optimized $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

D. Discussion on the parameter optimization procedure

Here we explain why both the BBA discounting operation using classifier weight $\boldsymbol{\alpha}$ and the BBA correction via confusion matrix $\boldsymbol{\beta}$ are included in the proposed optimal combination method. In discounting operation, the masses of belief on different classes for each classifier are proportionally discounted and the mass left is committed to the total ignorant element Ω according to the value of $\boldsymbol{\alpha}$. In fact, the discounting operation is used to control the influence of each classifier in the fusion by tuning the ignorance degree of each BBA. This is helpful to take fully advantage of the complementary information from different classifiers, and it can also reduce the harmful influence of the quite unreliable classifier that often produces errors. Nevertheless, the combination result of these discounted BBA's usually still has some discrepancy with the truth. Hence, the confusion matrix $\boldsymbol{\beta}$, which can be considered as the prior knowledge derived from the training data, is introduced to further modify the combination result by a judicious transformation of masses of belief among different classes. This BBA correction step can make the combination result as close as possible to the truth. Of course, if the combination result of the discounted BBA's has already been very close to the truth, then the confusion matrix $\boldsymbol{\beta}$ will be close to identity matrix. The discounting operation and BBA correction method work with quite distinct principles, and they are complementary to improve fusion performance. So both of them are necessary for achieving the best possible combination result.

In this proposed method, the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and the pattern weight $\mathbf{w}$ are iteratively optimized using the training data according to the minimization of criterion (11) under constraints (12). This is a normal constrained nonlinear least squares problem, and it can be solved by the active-set algorithm[8] [36]. The sequential quadratic programming (SQP) method can be used, and it solves a quadratic programming subproblem at each iteration. The estimate of the Hessian of the Lagrangian is updated at each iteration using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula (*i.e.* Quasi-Newton Method) [35], which always converges when the function has a quadratic Taylor expansion near an optimum. Once the optimized $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are determined, the pattern weight $\mathbf{w}$ is calculated by formula (13). Then $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be optimized again with the weighting vector. If the classification accuracy can be improved in this round of optimization, the weighting vector will be updated and the iterative optimization keeps going. Otherwise, we keep the optimized parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{w}$ in last optimization step, and the optimization procedure stops. Hence, the optimal value of $\mathbf{w}$ is determined depending on the improvement of accuracy.

IV. CAUTIOUS DECISION MAKING SUPPORT

In the applications, the class decision is often required for pattern classification according to the combination of classifiers. There exist many tools to deal with the uncertainty in decision making, such as probability, fuzzy sets [37], belief functions [38], [39], and so on. In the traditional way, the object is usually committed to the class with the biggest probability or fuzzy membership. In DST, a BBA is usually transferred into probability measure by pignistic probability

[8]In Matlab™software, the function *fmincon* is provided to solve such constrained nonlinear optimization problem.

transformation $BetP(.)$ [12] for decision making, and the pignistic probability of the singleton class ω_i is defined by

$$BetP(\omega_i) = \sum_{\substack{X \in 2^{\Omega} \\ \omega_i \in X}} \frac{1}{|X|} m(X). \tag{14}$$

The belief interval has been used in decision making under uncertainty. DST is incorporated in the modified version of the Analytic Hierarchy Process (AHP) [39], and it allows the numerical measures of uncertainty to be assigned to subsets of hypotheses as well as to individual hypothesis. The decision can be derived based on the belief interval as $[Bel(.), Pl(.)]$. In [38], DST has been also applied for the multi-attribute decision analysis with uncertainty, and the utility intervals is introduced to characterize the impact of ignorance due to the incompleteness in the assessment.

The decision maker usually wants to reach a specific decision (i.e. singleton class) for pattern classification. However, the hard decision often produces errors in the quite uncertain cases (e.g. several classes may take the close probabilities), and the error may yield dramatic consequences with important collateral damages in some applications like the target identification. In such case, the partially imprecise decision (i.e. set of several classes) must be preferable to a very prejudicial classification error. Nevertheless, how to balance the error and imprecision for pattern classification is not clearly addressed in previous works. So a cautious decision making strategy is introduced for the classification of uncertain data.

Let us consider an example to illustrate the problem. Suppose the combination result of classifiers for one pattern is $p(.) \triangleq BetP(.)$: $p(\omega_1) = 0.5$, $p(\omega_2) = 0.45$, and $p(\omega_3) = 0.05$. One sees that $p(\omega_2)$ is very close to $p(\omega_1)$, and it means that ω_1 and ω_2 appear undistinguishable for this pattern. If the pattern is classified to ω_1 by the hard decision making strategy, it will very likely cause an error. In such case, it could be better to cautiously commit the object to the set of classes $\{\omega_1, \omega_2\}$, because the partial imprecision is considered better than error. Moreover, the imprecision reminds the user that the available information is not sufficient for making a specific classification, and some other techniques should be included to make a clear decision. Nevertheless, the high imprecision of classification is not a good solution either. If one pattern can be classified to the singleton class with high confidence, it does not necessarily include any imprecision in decision. It seems interesting to find an efficient decision making strategy with a good compromise between imprecision and error.

An unified benefit value taking into account both the imprecision and error is presented here. Let us consider that one pattern belonging to a singleton class ω is classified to the set A containing either a singleton class or several classes. If $\{\omega\} \cap A = \emptyset$, it means this class decision is an error, and the benefit value of an error is considered as 0. If $A = \{\omega\}$, it is a correct decision, and the corresponding benefit value is given by 1. If $\omega \in A, |A| \geq 2$, it indicates that the real class is included in the decision set, but the decision is imprecise. The bigger cardinality value of $|A|$, the higher imprecision degree of decision. Of course, the high imprecision of classification produces the small benefit value. Hence, the benefit value $\mathcal{B}$ of the imprecision should be a monotone decreasing function of the cardinality value $|A|$, and it is simply defined by $(\frac{1}{|A|})^{\gamma}$ according to the above principle. The tuning parameter γ is the imprecision penalizing coefficient. The benefit value of the error, imprecision and correct classification can be defined by

$$\mathcal{B}_k(A) \triangleq \begin{cases} 0, & \text{if} \quad \{L(\mathbf{x}_k)\} \cap A = \emptyset. \\ (\frac{1}{|A|})^{\gamma}, & \text{if} \quad L(\mathbf{x}_k) \in A. \end{cases} \tag{15}$$

The real class label of $\mathbf{x}_k$ is denoted by $L(\mathbf{x}_k)$. The equality $\mathcal{B}_k(A) = 1$ holds if the correct decision is drawn as $A = \{L(\mathbf{x}_k)\}$ since $1^{\gamma} = 1$. So the benefit value of the imprecision and correct classification can be calculated by the common formula (the second part of eq. (15)). For a given imprecise decision set A, the benefit value will decrease when γ increases. It is argued that the benefit value obtained from the imprecise decision A (*i.e.* $(\frac{1}{|A|})^{\gamma}$) should be no less than that of random selection from A, such as $(\frac{1}{|A|})^{\gamma} > \frac{1}{|A|}$ (*i.e.* the probability of correct decision randomly selected in the set of A is equal to $\frac{1}{|A|}$). Thus, γ must be smaller than 1. Nevertheless, the benefit value of an imprecision classification must be smaller than a correct classification. Therefore $(\frac{1}{|A|})^{\gamma} < 1$, and one gets $\gamma > 0$. Hence, $\gamma \in (0, 1)$. In fact, the exact value of γ must be selected depending on the context of applications. If the error cost is rather large, one can choose a small γ value, and it implies that an imprecise classification is preferred to a classification error.

The expected decision strategy should make the total benefit value $\mathcal{B}_T$ as eq. (16) for the whole data set as big as possible.

$$\mathcal{B}_T = \sum_{k=1}^{K} \mathcal{B}_k(A). \tag{16}$$

One can see that the benefit value defined in eq. (16) is closely related with the set A.

In this work, a simple decision criteria is adopted, and the pattern will be committed to a class set A defined by $A = \{\omega_i | p(\omega_i) \geq \epsilon \cdot \max\{p(.)\}\}$ with the threshold $\epsilon \in (0, 1]$. The class set A consists of classes having a probability close to the maximum one in the classification result with respect to the threshold ϵ. For each class $\omega_i, i = 1, \ldots, c$, the parameter ϵ_i corresponding to the maximum benefit value may be different. Hence, the different optimal values of ϵ_i, $i = 1, \ldots, c$ will be found to maximize the total benefit value defined in (16) using each class of training data by a grid-search method[9].

This cautious decision making strategy is chosen mainly to draw a decision from the soft output of ensemble classifier. The decision is just a binary value, and it cannot reflect so much useful classification knowledge as the original soft output of ensemble classifier. The decision making strategy is generally not directly related with the design of (ensemble) classifier. In this proposed method, we want the combination result of classifiers as close as possible to the truth. Hence, the parameters are obtained by minimizing the bias of the soft combination results with respect to the ground truth. This is a very often used optimization strategy to minimize the

[9] A proper interval of $\epsilon \in [0.5, 1]$ is recommended here.

discrepancy between the system output and the expected value in the classification problem [4], [16], [27], [31], and the decision making is not involved in the parameter optimization procedure. We also adopt such normal optimization way here. It is worth noting that this cautious decision-making strategy is very general and it can be used in all applications where a decision must be made from the soft probabilistic output.

V. Experiment applications

The classification performance of this new method called Optimal Weighted DS (OWDS) combination rule will be evaluated and compared with several other fusion methods, such as simple and weighted averaging rule, simple and weighted DS combination rule. The weight of each classifier is usually determined according to the classification accuracy AC, and the individual accuracy of the classifier $\mathcal{C}_l$ is $AC_l, l = 1, \ldots, n$ as $AC_l \triangleq \frac{N_l}{N_T}$, where the number of patterns correctly classified by $\mathcal{C}_l$ is N_l, and the number of total patterns is N_T. The commonly used classifier weight say $\alpha_l \in [0, 1]$ can be calculated by $\alpha_{l1} = AC_l$ or $\alpha_{l2} = \frac{AC_l - AC_L}{AC_U - AC_L}$ where $AC_U = \max_l AC_l$ and $AC_L = \min_l AC_l$. The normal hard decision making strategy is used to calculate the accuracy, and the object is assigned to the class with the maximum probability.

For the simple weighted average combination rule, the sum of normalized weighting factors of classifiers must be equal to one. In DS combination, one does not need to consider normalized weighting factors because the discounting is done separately on the BBA output of each classifier. We can directly use $\alpha_{l1} \in [0, 1]$ or $\alpha_{l2} \in [0, 1]$ as the weighting factor.

Five related fusion methods have been evaluated in this work: 1) the simple Average Fusion (AF); 2) the Weighted Average Fusion (WAF); 3) the Average Fusion with Learning of Weight (AFLW); 4) Dempster's fusion rule (DS); and 5) the Weighted DS fusion rule (WDS). The brief description of these methods is shown in Table II.

Table II
Description of the used fusion methods.

Name	Calculation
AF	$\mathbf{p} = \frac{1}{n} \sum_{l=1}^{n} \mathbf{p}_l.$
WAF	$\mathbf{p} = \sum_{l=1}^{n} \tilde{\alpha}_l \mathbf{p}_l.$
AFLW	$\mathbf{p} = \sum_{l=1}^{n} \hat{\alpha}_l \mathbf{p}_l.$
DS	$\mathbf{m} = \mathbf{m}_1 \oplus \ldots \oplus \mathbf{m}_n.$
WDS	$\mathbf{m} = {}^{\alpha_1}\mathbf{m}_1 \oplus \ldots \oplus {}^{\alpha_n}\mathbf{m}_n.$

In Table II, the meaning of the symbol is given by: $\mathbf{p}_l$ being probabilistic output of classifier $\mathcal{C}_l$, $\tilde{\alpha}_l$ being the normalized weighting factors, $\hat{\alpha}$ being the optimal weighting factor learned by minimizing the distance between the weighted averaging combination result and the ground truth as done in [4]. Both α_{l1} and α_{l2} will be used to calculate the classifier weight in WAF and WDS rules, and the higher classification accuracy is reported in following Tables IV–IX.

The base classifier can be selected according to the actual applications. In this work, Support Vector Machine (SVM) [42], naive Bayesian Classifier (BC) [43] and Evidential Neural Network (ENN) [16] classifier are employed as the base classifiers. In SVM, we use the one class versus the others classification strategy, and the normal linear kernel is adopted as $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$. The classifier output has been transferred to the probability measure in order to preserve the useful classification information as most as possible in the combination procedure. In a c-class problem, the output of SVM classifier for object $\mathbf{y}$ is denoted by $\mathbf{f} = (f_1, f_2, \ldots, f_c)$, and f_i represents the hyperplane function value of SVM for class ω_i versus the other classes. The transferred probability is defined by $\mathbf{p} = (p_1, p_2, \ldots, p_c)$ with $p_i = \frac{f_i - \min_j f_j}{\sum_{g=1}^{c} (f_g - \min_j f_j)}$. This transformation is similar to the max-min normalization procedure, and thus the bigger hyperplane function value in output corresponds to bigger probability value. The transferred probabilities also satisfies the condition $\sum_{i=1}^{c} p_i = 1$, $p_i \in [0, 1]$. Of course, some other kernels and other probability transformation methods can be selected according to the actual application.

The BBA output of ENN consists of the singletons and the total ignorance. The output of Bayesian classifier is a probability measure. Both the BBA and probability can be directly used in the proposed optimal combination of classifiers. The base classifier(s) will be respectively trained using different subsets of attributes, and the multiple classification results obtained by different classifiers will be combined for classifying the objects. The pignistic probability transformation $BetP(.)$ is used to transform a BBA into a probability measure for making a decision. The hard decision-making approach and the new cautious decision-making approach are both applied and evaluated with the classifier fusion methods.

Twelve real data sets from UCI repository (http://archive.ics.uci.edu/ml) have been used here to evaluate the performances of our new OWDS method with respect to the other methods. Each data set includes one or two cases (*i.e.* the attribute set is divided into different subsets for different classifiers). Hence, there are total twelve real data sets consisting of twenty cases in the experiments. The basic knowledge of the used data sets is shown by Table III. For each data set, the patterns consist of multiple attributes, and these attributes will be randomly divided into n distinct subsets without overlapping attributes, and each subset of attributes will be respectively used to train the base classifier (SVM, ENN and BC). The k-fold cross validation is often used for the classification performance evaluation, but k remains a free parameter. We use the simple 2-fold cross validation here, since the training and test sets are large, and each sample can be respectively used for training and testing on each fold. For each fold, the program is randomly run ten times. The average classification accuracy and benefit values with the standard deviation for different methods are reported in Tables IV–IX.

In the Tables IV–IX and figures 1–2, OWDS corresponds to the proposed optimal weighted DS combination method where each pattern is considered with same importance (*i.e.*

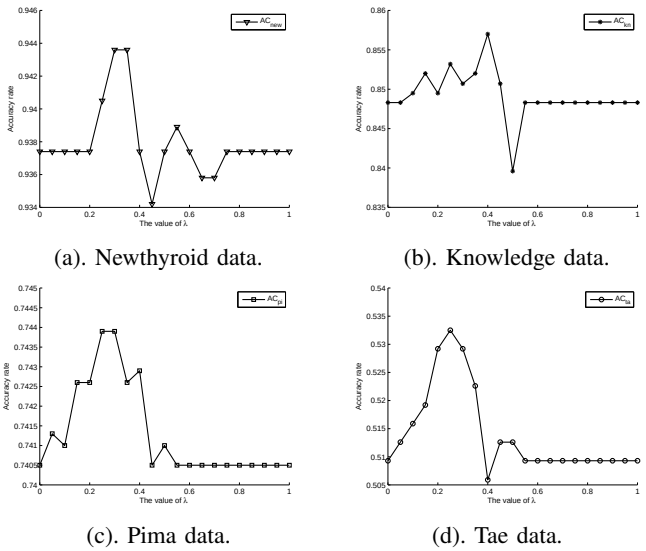
all the training patterns have the same weight). OWDS-PW corresponds to the Optimal Weighted DS combination method where the Pattern Weight (PW) is automatically tuned using the proposed method. The benefit value $\mathcal{B}_T$ (defined in eq.(16)) for all the fusion methods on different data sets based on the cautious decision making strategy is also reported for SVM classifier and for the hybrid classifier (based on a random selection of SVM, ENN and BC) in Tables V and IX. The lower and upper accuracy (given by the average value over multiple runs) of the singleton classifiers to combine are respectively denoted by $AC_L = \min_l AC_l$ and $AC_U = \max_l AC_l$, where AC_l for $l = 1, \ldots, n$ is the classification accuracy of the individual classifier $\mathcal{C}_l$. The average of accuracy (or benefit value) denoted by Ave on different data sets with the same fusion method is given in the second last row of the tables V–IX to show the general performance of the method. Moreover, Winning Times[10] (denoted by WT) of each fusion method on the twenty classification cases is also reported in the last row of Tables V–IX.

The influence of the tuning of parameters λ and γ on the classification result is evaluated at first in experiment 1, and then the performance of different fusion methods are evaluated and compared with different base classifiers in experiment 2.

A. Experiment 1: Test of parameter influence on fusion performance

There are two parameters involved in the proposed method, *i.e.* the distance penalizing coefficient λ associated with the pattern weight as given in eq.(13), and the imprecision penalizing coefficient γ in eq.(15) for cautious decision making. In this experiment we evaluate their influence on the classification performance.

We take the following four real data sets from UCI to show the parameter influence: 1) newthyroid; 2) knowledge; 3) pima; and 4) tae data sets. The attributes of each data set is randomly divided into two different subsets for different classifiers, and three base classifiers, *i.e.* SVM, ENN and BC are randomly selected for each subset of attribute in a data set. The classification accuracy (*i.e.* average value for ten-times running) of the proposed method OWDS-PW with the tuning of parameter λ is shown in Fig. 1 for different data sets.



(a). Newthyroid data. (b). Knowledge data.

(c). Pima data. (d). Tae data.

Figure 1. Classification accuracy on several data sets with different λ.

One sees that the accuracy changes with the tuning of λ value, and the optimal λ values in different data sets are different. Hence, it is difficult to give a common optimal value for different applications. We can use the training data set to seek the optimal λ value in each application, and the optimal value should correspond to the highest accuracy. Nevertheless, this optimization procedure could be time-consuming. One observes in Fig. 1 that the proposed method generally produces good performance if $\lambda \in (0.2, 0.4)$. We find the high accuracy usually can be reached when one takes $\lambda = 0.25$ and that is why we recommend $\lambda = 0.25$ as the default value for λ. In the following experiments, we have used this default value.

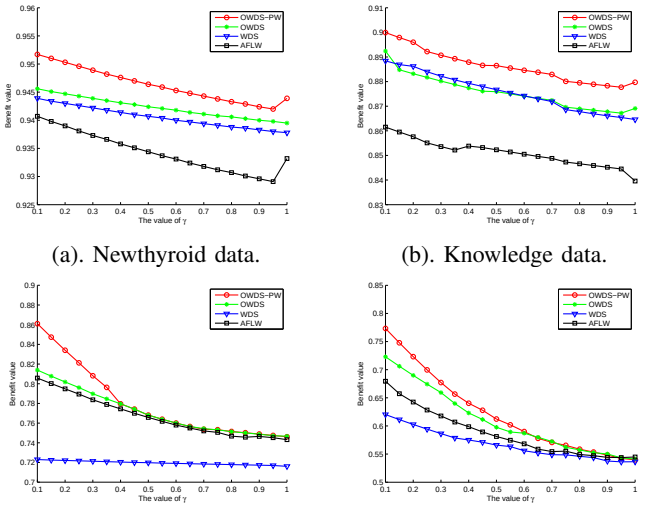
We have also tested the influence of tuning of γ on the benefit value for cautious decision making with different data sets. The change curves of benefit values of different methods including OWDS, OWDS-PW, AFLW and WDS are shown in Fig. 2.

We see that the benefit value of the four methods generally decreases with the decreasing of γ value. This is reasonable behavior because the smaller γ value yields the bigger benefit value for the imprecise decision. The determination of γ value mainly depends on the application. If the error cost is quite large, then an imprecise classification decision must be preferred to an error, and one can take the smaller γ value. If the error cost is not very large, then one should take the big γ value (*i.e.* close to one). We take $\gamma = 0.8$ in the following experiments to test the performance of our proposed method with respect to several other related methods, and the

[10]If one method produces the maximum accuracy/benefit value for one classification case compared with the other fusion methods, it wins one time. Several different fusion methods may produce the same maximum accuracy/benefit value, and they are all considered winner in such case.

Table III
BASIC INFORMATION OF THE USED DATA SETS.

Data	Class	Attribute	Instance
newthyroid (new)	3	4	215
white Wine quality (wq)	7	11	4898
knowledge(kn)	4	5	403
Wbdc (Wb)	2	30	569
red wine quality (rwq)	6	11	1599
pima(pi)	2	8	768
tae(ta)	3	5	151
satimage (sat)	7	36	6435
magic (ma)	2	10	19020
vehicle (ve)	4	18	946
page-blocks (pb)	5	10	5472
texture (te)	11	40	5500



(a). Newthyroid data. (b). Knowledge data.

(c). Pima data. (d). Tae data.

Figure 2. Benefit value of several methods with different γ.

imprecise classification decision strategy is preferred to the random selection (corresponding to $\gamma = 1$) in decision making.

B. Experiment 2: Classification with different base classifiers

In this experiment, three base classifiers including SVM, ENN and BC are used to test the performance of the proposed fusion method with respect to the other related methods. The classification accuracy (*i.e.* hard decision) of different fusion methods with the singleton base classifier (*i.e.* SVM, ENN or BC) is respectively given in Tables IV, VI and VII. The common base classifier is operated on the different attribute subsets for each data set. The hybrid base classifier fusion is also tested, where the three base classifiers SVM, ENN and BC are randomly selected for classifying each subset of attribute in a data set. For example, the Wbdc data set consists of 30 attributes that are randomly divided into 3 subsets. Then, one subset of attribute is classified by SVM, and another one is classified by ENN, and the last one is classified by BC base classifier. Then their classification results are combined by the different methods, and the classification accuracy according to the combination results is reported in Table VIII. In the cautious decision making strategy, the benefit value of different fusion methods with SVM base classifier and the hybrid base classifier is respectively given in tables V and IX. In the following tables, the maximum of accuracy and benefit value is emphasized in boldface for convenience.

In the experiment 2, one sees that the proposed classifier fusion methods (*i.e.* OWDS and OWDS-PW) with the optimal classifier weight and confusion matrix generally produce higher accuracy than other methods in most cases (according to the average accuracy on different data sets *Ave* and the winning times *WT*) as shown in last rows of Tables IV–VIII. This is because the classifier weight is determined by globally optimizing the fusion result as well as the confusion matrix. In other methods, the classifier weight is calculated according to the accuracy of individual classifiers, and the complementary knowledge of different classifiers is not efficiently taken into account. The training pattern weight is considered equal in the traditional classifier weight determination methods. In the new OWDS-PW method, the patterns hard to classify play a more important role in the parameter optimization, and the pattern weight is automatically tuned for obtaining the best possible classification results. We find that the accuracy can be further improved when the optimal (rather than equal) pattern weight is employed in OWDS-PW with respect to OWDS according to the Tables IV–VIII. Nevertheless, we also find that AFLW and several other methods can produce a bit better performance (higher accuracy and benefit value) than our proposed method OWDS-PW in some cases (according to the winning times *WT*). This is because AFLW and OWDS-PW work with different combination rules. OWDS-PW working with DS combination rule is suitable for dealing with the independent and complementary sources of information, and it can produce good performance taking advantage of the complementarity of classifiers. AFLW can well handle the random cases to obtain the average value. The performance of OWDS-PW may be not as good as AFLW when the classification results provided by different classifiers are not very complementary.

The accuracy is calculated based on the hard decision that the object is assigned to the class with maximum probability. If the proposed cautious decision making support strategy is applied, the benefit value shown in Tables V and IX is usually bigger than the accuracy value, and it implies that the error has been reduced by the cautious decision making strategy. This is because the partial imprecision is reasonably kept in the cautious decision. It is considered that the imprecision is preferred to error, and the imprecision gains bigger benefit value than error. The partial imprecision can also warn the user that the used knowledge is not sufficient for the specific classification of pattern, and some other sources of information are essential for making more specific (refined) classification. The proposed method OWDS-PW produces bigger benefit value than the other methods in most cases as shown in Tables V and IX. This shows the effectiveness and potential interest of this new method.

Nevertheless, the proposed method has bigger computation complexity compared with the other related methods due to the optimization of the classifier weight, confusion matrix and the cautious decision threshold. Fortunately, these optimization procedures can be done off-line using the training data, and it can be easily implemented with some mathematical software like Matlab^{TM}. Generally speaking, the computation complexity of the proposed method is the *price to pay* for improving the classification accuracy. In our future works, we will try to improve the calculation efficiency especially for dealing with large data sets using new techniques, like random sampling.

VI. Conclusion

We have proposed a new weighted combination method for multiple classifiers based on evidential reasoning. The weighting factors of classifiers are globally optimized by minimizing the error criteria, which is defined by the distance between the combination result of classifiers and the target value (*i.e.* truth of classification) in training data space. In

order to achieve the best classification performance, a confusion matrix is also introduced to characterize the probability of the object belonging to one class but classified into another class according to combination result. This matrix is used to further modify the combination result for making it as close as possible to the target value, and it is optimized using the training data as well as the classifier weight. Moreover, the training patterns hard to classify are considered playing a more important role in the parameter optimization than the patterns easy to classify. The pattern weight is automatically tuned according to the distance between classification result and the truth, and the bigger distance generally leads to the bigger weight. A cautious decision making method has been also presented. The partial imprecision is introduced to reduce the error cost, because imprecise classification decision is preferred to error. Various real data sets have been used to test the performance of the new method, and our results and analysis show that the new method can efficiently improve the accuracy of the classification and provide a higher unified benefit value than other related methods in most cases.

Acknowledgements

This work has been partially supported by National Natural Science Foundation of China (Nos.61672431, 61403310) and the Fundamental Research Funds for the Central Universities – China (No.3102017zy020).

REFERENCES

[1] Z.-H. Zhou, J. Wu, W. Tang, *Ensembling neural networks: Many could be better than all*, Artif. Intell., Vol. 137(1-2), pp. 239–263, 2002.

[2] L.I. Kuncheva, *A Theoretical Study on Six Classifier Fusion Strategies*, IEEE Trans. on Pattern Anal. Mach. Intell., Vol. 24(2), 2002.

[3] D. Ruta, B. Gabrys, *An Overview of Classifier Fusion Methods*, Computing and Information Systems, Vol. 7, pp. 1–10, 2000.

[4] J. Yang, X. Zeng, S. Zhong, *Effective Neural Network Ensemble Approach for Improving Generalization Performance*, IEEE Trans. Neural Networks and Learning Systems, Vol. 24 (6), pp. 878–887, 2013.

[5] L. Kuncheva, J. Bezdek, R. Duin, *Decision templates for multiple classifier fusion: an experimental comparison*, Pattern Recognition, Vol. 34(2), pp. 299–314, 2001.

[6] N.J. Pizzi, W. Pedrycz, *Aggregating multiple classification results using fuzzy integration and stochastic feature selection*, International Journal of Approximate Reasoning, Vol. 51(8), pp. 883–894, 2010.

[7] B. Quost, M.-H. Masson, T. Denœux. *Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules*, International Journal of Approximate Reasoning, Vol. 52(3), pp. 353–374, 2011.

[8] Y.X. Bi, J.W. Guan, D. Bell, *The combination of multiple classifiers using an evidential reasoning approach*, Artificial Intelligence, Vol. 172, pp. 1731–1751, 2008.

[9] J.B. Yang, D.l. Xu, *Evidential reasoning rule for evidence combination*, Artificial Intelligence, Vol. 205, pp. 1–29, 2013.

[10] G. Shafer, *A mathematical theory of evidence*, Princeton Univ. Press, 1976.

[11] F. Smarandache, J. Dezert (Editors), *Advances and applications of DSmT for information fusion*, American Research Press, Rehoboth, Vol. 1-4, 2004–2015.

[12] P. Smets, *Decision making in the TBM: the necessity of the pignistic transformation*, International Journal of Approximate Reasoning, Vol. 38, pp. 133–147, 2005.

[13] A.-L. Jousselme, D. Grenier, É. Bossé, *A new distance between two bodies of evidence*, Information Fusion, Vol. 2, No. 2, pp. 91–101, 2001.

[14] A.-L. Jousselme, C. Liu, D. Grenier, É. Bossée, *Measuring ambiguity in the evidence theory*, IEEE Trans. on Systems, Man & Cybernetics, Part A: Systems, Vol. 36(5), pp. 890–903, 2006.

[15] S. Destercke, P. Buche, B. Charnomordic, *Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse*, IEEE Trans. on Knowl. Data Eng., Vol. 25(1), pp. 92–105, 2013.

[16] T. Denœux, *A neural network classifier based on Dempster-Shafer theory*, IEEE Trans. on Systems, Man & Cybernetics A, Vol. 30, No. 2, pp. 131–150, 2000.

[17] Z.-g. Liu, Q. Pan, J. Dezert, G. Mercier, *Credal classification rule for uncertain data based on belief functions*, Pattern Recognition, Vol. 47, No. 7, pp. 2532–2541, 2014.

[18] Z.-g. Liu, Q. Pan, G. Mercier, J. Dezert, *A new incomplete pattern classification method based on evidential reasoning*, IEEE Trans. on Cybernetics, Vol. 45, No. 4, pp. 635–646, 2015.

[19] C. Lian, S. Ruan, T. Denœux, *Dissimilarity metric learning in the belief function framework*, IEEE Trans. on Fuzzy Systems, Vol. 24(6):1555–1564, 2016.

[20] T. Denœux, P. Smets, *Classification using belief functions: relationship between case-based and model-based approaches*, IEEE Trans. on Systems, Man & Cybernetics, Part B: Vol. 36, No. 6, pp. 1395–1406, 2006.

[21] T. Denoeux, S. Sriboonchitta, O. Kanjanatarakul, *Evidential clustering of large dissimilarity data*, Knowledge-Based Systems, Vol. 106, pp. 179–195, 2016.

[22] T. Denœux, *A k-nearest neighbor classification rule based on Dempster-Shafer Theory*, IEEE Trans. on Systems, Man & Cybernetics, Vol. 25, No. 5, pp. 804–813, 1995.

[23] E. Ramasso, T. Denœux, *Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions*, IEEE Trans. on Fuzzy Systems, Vol. 22(2), pp. 395–405, 2014.

[24] Z.j Zhou, C.h Hu, G.y Hu, X.x Han, B.c Zhang, Y.w Chen,*Hidden Behavior Prediction of Complex Systems Under Testing Influence Based on Semiquantitative Information and Belief Rule Base*, IEEE Trans. on Fuzzy Systems, Vol. 23(6), pp. 2371–2386, 2015.

[25] X. Deng, Y. Hu, F. T. S. Chan, S. Mahadevan, Y. Deng, *Parameter estimation based on interval-valued belief structures*, European Journal of Operational Research, Vol. 241(2), pp. 579–582, 2015.

[26] T. Denœux, *Maximum likelihood estimation from uncertain data in the belief function framework*, IEEE Trans. on Knowledge and Data Engineering, Vol. 25, No. 1, pp.119–130, 2013.

[27] D. Mercier, B. Quost, T. Denœux, *Refined modeling of sensor reliability in the belief function framework using contextual discounting*, Information Fusion, Vol. 9(2), pp. 246–258, 2008.

[28] D. Mercier, G. Cron, T. Denœux, M.H. Masson, *Decision fusion for postal address recognition using belief functions*, Expert Systems with Applications, Vol. 36(3), pp. 5643–5653, 2009.

[29] F. Pichon, D. Mercier, E. Lefevre , F. Delmotte, *Proposition and learning of some belief function contextual correction mechanisms*, International Journal of Approximate Reasoning, Vol. 72, pp. 4–42, 2016.

[30] A. Martin, E. Radoi, *ATR algorithms using information fusion models*, in Proc. of 7th International Conference on Information Fusion (FUSION 2004), Stockholm, Sweden, July 2004.

[31] A. Al-Ani, M. Deriche, *A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence*, J. Artif. Intell. Res., Vol. 17, pp. 333–361, 2002.

[32] F. Moreno-Seco, J.M. Inesta, P.J. Ponce de Leon, L. Mico, *Comparison of Classifier Fusion Methods for Classification in Pattern Recognition Tasks*, D.-Y. Yeung et al. (Eds.): Springer-Verlag Berlin Heidelberg, pp. 705–713, 2006.

[33] E. Lefevre, O. Colot, P. Vannoorenberghe, *Belief function combination and conflict management*, Information Fusion, Vol. 3(2), pp. 149–162, 2002.

[34] J. Dezert, A. Tchamova, *On the validity of Dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule*, International Journal of Intelligent Systems, Vol. 29, No. 3, pp. 223–252, March 2014.

[35] J. Nocedal, S.J. Wright, *Numerical Optimization (2nd ed.)*, Berlin, New York: Springer-Verlag, 2006.

[36] S.P. Han,*A Globally Convergent Method for Nonlinear Programming*, Journal of Optimization Theory and Applications, Vol. 22, p. 297, 1977.

[37] R. M. Rodriguez , L. Martinez ,F. Herrera, *Hesitant Fuzzy Linguistic Term Sets for Decision Making*,IEEE Trans. on Fuzzy Systems, Vol. 20(1), pp. 109–119, 2012.

[38] J.B. Yang, D.L. Xu, *On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty*, IEEE Trans. on Systems, Man, and Cybernetics, Part A, Vol. 32(3), pp. 289–304, 2002.

[39] M. Beynon, B. Curry, P. Morgan, *The Dempster-Shafer theory of evidence: an alternative approach to multi-criteria decision modeling*, Omega, Vol. 28, pp. 37–50, 2000.
[40] A. Jousselme, P. Maupin, *Distances in evidence theory: Comprehensive survey and generalizations*, Int. J. of Approx. Reasoning, Vol. 53(2), pp.118–145, 2012.
[41] A. Essaid, A. Martin, G. Smits, B. Ben Yaghlane, *A distance-based decision in the credal level*, International Conference on Artificial Intelligence and Symbolic Computation, Sevilla, Spain, 2014.
[42] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[43] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall (second edition), 2003.

Zhun-ga Liu was born in China, 1984. He received the Bachelor, Master and Ph.D degree from Northwestern Polytechnical university (NPU), Xi'an, China in 2007, 2010 and 2013 respectively. He is an associate professor in School of Automation, NPU. His research interest mainly focuses on belief function theory and its application in pattern recognition.

Quan Pan was born in China, 1961. He received the Bachelor degree in Huazhong University of Science and Technology, Wuhan, China, and he received the master and doctor degree in Northwestern Polytechnical University (NPU), Xi'an, China in 1991 and 1997. He has been professor since 1998 in NPU, and he has been dean of School of Automation, NPU since 2009. His main research interests are information fusion and pattern recognition.

Jean Dezert was born in France, 1962. He received the Electrical Engineering degree in 1985, and his Ph.D degree from the University Paris XI, 1990. Since 1993, he has been senior research scientist at the French Aerospace Lab. His main research interest focuses on decision-making support and belief function theory.

Arnaud Martin was born in France, 1974. He received the Master and Ph.D degree from University of Rennes 1, France in 1998 and 2001 respectively. He has been professor in at University of Rennes 1, IUT of Lannion since 2010. His research interest mainly focuses on belief function theory with the applications.

Table IV
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH SVM BASE CLASSIFIER (IN %).

Data	n	$[AC_L,\ AC_U]$	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	[89.30, 90.32]	90.89±1.90	91.07±1.62	91.16±1.13	90.89±1.90	90.79±1.81	92.37±0.41	**93.12±0.76**
wq	5	[45.13, 46.76]	45.51±0.07	45.19±0.25	49.26±0.62	45.49±0.04	45.45±0.11	51.09±0.75	**51.57±0.15**
wq	2	[46.38, 49.69]	49.20±1.64	49.57±3.34	49.81±1.62	48.98±1.33	49.67±2.75	51.03±0.07	**51.35±1.21**
kn	2	[49.38, 56.94]	45.10±1.27	52.11±7.65	60.18±5.85	45.75±1.25	55.22±6.88	58.21±5.37	**60.60±5.94**
Wb	6	[89.61, 91.28]	93.48±0.96	90.14±2.49	91.85±2.78	87.79±2.52	92.98±1.40	**93.90±1.42**	**93.90±1.42**
Wb	3	[91.03, 91.60]	92.95±1.44	91.26±4.14	91.95±4.70	91.74±2.07	93.18±1.82	**93.30±1.66**	**93.30±1.66**
rwq	5	[47.05, 51.16]	55.24±0.69	52.39±1.32	53.47±0.78	55.56±0.57	55.41±1.10	56.29±0.62	**56.97±0.55**
rwq	2	[52.62, 53.10]	55.72±1.82	55.31±2.91	55.83±2.08	56.00±1.75	55.77±2.32	56.38±1.37	**56.95±0.92**
pi	4	[65.31, 74.74]	66.56±0.41	70.83±5.30	**74.87±4.65**	65.29±0.27	69.69±2.00	74.74±0.18	74.82±0.41
pi	2	[67.14, 74.71]	68.88±0.95	69.64±4.16	**74.82±2.46**	68.88±0.95	69.64±4.16	**74.82±2.46**	**74.82±2.46**
ta	2	[45.70, 46.23]	49.00±3.28	48.88±3.77	49.68±3.68	47.56±4.15	49.94±3.14	51.17±2.78	**51.26±2.05**
sat	6	[50.82, 70.51]	75.75±0.86	73.87±0.50	72.91±1.19	75.69±1.00	75.75±0.56	75.87±0.03	**76.16±0.82**
sat	3	[71.97, 73.94]	76.24±0.48	75.29±0.57	74.32±1.91	**76.35±0.41**	75.83±0.60	76.07±0.37	76.22±0.55
ma	5	[64.84, 74.60]	69.73±1.87	68.12±4.64	72.45±0.82	69.73±1.87	70.16±2.48	**74.97±0.29**	**74.97±0.29**
ma	2	[70.45, 78.87]	72.05±0.04	**78.87±0.37**	**78.87±0.37**	72.92±0.23	**78.87±0.37**	**78.87±0.37**	**78.87±0.37**
ve	6	[35.52, 50.30]	51.60±7.77	51.89±0.17	51.73±1.92	44.50±2.59	53.96±6.94	**56.65±3.18**	56.50±3.01
ve	9	[32.21, 47.52]	47.58±3.59	48.58±0.50	50.74±4.18	36.11±1.42	48.17±3.43	51.65±1.84	**52.25±1.00**
pb	5	[89.76, 90.57]	89.91±0.12	90.33±0.53	90.50±0.17	90.27±0.63	90.25±0.58	90.12±0.48	**90.70±1.32**
te	8	[69.37, 84.97]	93.68±0.35	90.95±1.14	72.75±2.21	**93.85±0.28**	93.67±0.41	**93.85±0.28**	**93.85±0.28**
te	4	[89.96, 94.04]	96.79±0.42	95.96±1.01	86.75±2.68	**96.81±0.42**	96.75±0.34	96.78±0.47	**96.81±0.42**
Ave		[63.18, 63.18]	69.29±1.50	69.51±2.32	69.70±2.29	68.01±1.28	70.56±2.16	72.41±1.22	**72.75±1.28**
WT			0	1	3	3	1	7	17

Table V
BENEFIT VALUE OF DIFFERENT COMBINATION METHODS WITH SVM BASE CLASSIFIER (IN %).

Data	n	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	91.12±1.18	91.38±1.10	91.28±1.07	91.92±1.07	90.72±1.14	92.71±0.21	**93.29±0.59**
wq	5	49.87±0.08	49.98±0.96	50.45±0.07	46.57±0.17	49.91±0.11	52.24±0.10	**52.34±0.30**
wq	2	50.87±1.46	50.51±2.03	51.83±0.99	50.89±0.94	50.78±0.94	52.02±0.43	**52.34±0.92**
kn	2	51.05±0.95	58.71±4.38	**68.76±5.62**	50.04±1.75	61.89±5.23	63.69±4.63	66.54±5.35
Wb	6	93.38±0.57	90.14±2.49	92.77±2.35	85.27±2.74	93.23±1.86	93.90±1.42	**94.01±1.36**
Wb	3	93.28±1.73	92.05±5.04	92.86±4.65	92.11±3.52	93.15±1.76	**93.33±1.75**	93.29±1.56
rwq	5	57.02±0.85	53.65±1.57	55.77±0.75	57.23±0.36	57.68±1.05	58.55±0.92	**58.79±0.46**
rwq	2	55.88±1.97	55.71±3.01	56.29±2.13	56.71±1.59	56.92±2.13	57.70±1.25	**58.26±1.10**
pi	4	70.27±1.31	72.51±4.35	**75.02±3.79**	66.31±0.65	71.14±2.33	74.74±0.18	74.78±0.42
pi	2	73.34±0.89	73.88±3.05	**74.82±2.46**	73.02±1.21	74.13±3.75	**74.82±2.46**	**74.82±2.46**
ta	2	52.26±1.95	50.46±1.60	52.34±2.69	49.18±5.95	52.11±2.78	**52.67±2.78**	52.52±2.18
sat	6	76.22±0.61	74.29±1.19	73.70±1.38	75.81±0.75	76.59±0.61	76.37±0.40	**76.89±0.57**
sat	3	76.57±0.07	76.42±0.10	75.05±2.05	76.67±0.05	76.60±0.40	76.60±0.40	**77.04±0.52**
ma	5	71.50±1.11	70.05±3.48	73.86±0.70	69.73±1.87	70.16±2.48	75.11±0.41	**75.91±1.04**
ma	2	75.94±0.29	**78.87±0.37**	**78.87±0.37**	75.67±0.33	**78.87±0.37**	**78.87±0.37**	**78.87±0.37**
ve	6	52.41±5.60	53.83±1.07	53.47±1.03	46.96±1.58	54.68±3.55	58.48±3.23	**59.71±0.75**
ve	9	51.38±2.31	50.41±0.76	53.95±3.44	47.94±5.03	53.06±1.21	53.52±1.12	**55.43±2.51**
pb	5	90.59±0.15	91.09±0.66	91.12±0.32	90.41±0.83	90.65±0.75	91.21±0.53	**92.47±0.59**
te	8	93.91±0.89	91.73±1.12	75.56±2.00	93.94±0.58	93.84±0.62	93.94±0.58	**94.15±0.25**
te	4	96.81±0.38	96.08±1.00	87.90±1.75	**96.96±0.22**	96.94±0.18	96.85±0.41	96.85±0.41
Ave		71.18±1.22	71.09±1.97	71.28±1.99	69.67±1.56	72.15±1.64	73.37±1.18	**73.92±1.19**
WT			1	4	1	1	4	15

Table VI
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH ENN BASE CLASSIFIER (IN %).

Data	n	$[AC_L,\ AC_U]$	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	[86.79, 88.85]	92.10±0.57	91.26±1.79	92.50±0.32	92.06±0.70	92.01±1.01	92.81±0.33	**93.22±0.25**
wq	5	[45.22, 46.82]	45.75±0.90	46.07±1.09	48.97±0.25	45.62±0.77	45.84±0.87	48.80±2.95	**49.67±3.10**
wq	2	[45.27, 47.35]	47.24±1.90	46.62±1.17	46.55±1.75	46.86±1.25	46.92±1.44	47.22±2.92	**47.71±2.04**
kn	2	[34.05, 74.50]	70.60±5.57	70.15±6.21	72.76±5.98	67.21±4.99	73.87±6.17	74.93±5.35	**75.95±5.46**
Wb	6	[82.70, 87.57]	91.23±0.68	90.90±1.92	**93.33±1.99**	91.35±0.75	91.24±0.79	92.82±1.61	93.21±0.83
Wb	3	[83.64, 87.25]	90.63±0.78	90.47±1.15	91.53±0.91	90.73±0.92	90.70±1.03	90.86±0.99	**91.93±1.11**
rwq	5	[42.02, 51.57]	56.62±0.63	54.64±0.59	57.12±0.69	56.81±0.82	56.47±0.88	57.47±0.62	**58.31±0.84**
rwq	2	[47.82, 54.08]	55.31±4.14	54.26±3.93	54.15±4.22	54.87±3.86	55.01±4.03	55.11±3.52	**55.62±4.26**
pi	4	[64.97, 74.53]	70.21±1.01	70.05±3.09	74.92±3.31	72.03±0.92	72.53±1.25	**76.04±1.10**	75.98±0.09
pi	2	[65.10, 70.31]	70.68±4.97	69.17±3.62	70.72±4.28	70.63±4.76	70.73±4.94	71.46±4.33	**71.72±4.65**
ta	2	[38.17, 39.96]	45.02±3.45	43.04±4.04	43.03±3.76	45.46±4.71	44.58±4.42	**46.78±2.99**	**46.78±2.99**
sat	6	[70.63, 79.83]	82.86±0.51	79.11±0.14	82.87±0.42	82.56±0.44	82.72±0.57	83.18±0.45	**83.59±0.48**
sat	3	[77.47, 80.92]	82.51±0.06	81.86±0.49	82.60±0.87	82.18±0.12	82.29±0.21	82.48±0.01	**82.89±0.07**
ma	5	[64.84, 74.82]	70.91±0.07	73.20±4.13	72.99±2.97	71.40±0.39	71.54±0.50	77.17±0.16	**77.44±0.13**
ma	2	[64.83, 68.07]	67.45±3.68	67.67±4.01	67.74±3.24	67.56±3.85	67.58±3.87	67.78±4.73	**67.99±4.46**
ve	6	[37.77, 50.77]	56.32±2.59	53.37±0.92	56.09±2.93	57.45±1.84	57.39±1.59	58.33±1.09	**58.63±0.84**
ve	9	[34.69, 49.76]	56.15±1.84	50.47±4.01	58.46±4.47	55.97±1.76	55.73±1.59	59.63±2.59	**60.34±2.26**
pb	5	[89.77, 91.16]	89.77±0.00	89.77±0.00	**92.51±1.34**	89.77±0.00	89.77±0.00	91.08±0.03	91.08±0.03
te	8	[56.09, 72.39]	81.21±1.71	77.89±2.78	81.15±2.79	81.00±2.19	81.35±2.29	82.20±1.65	**82.66±1.38**
te	4	[67.76, 74.31]	77.55±1.27	76.48±0.35	78.65±1.21	77.46±0.91	77.57±0.96	78.83±0.42	**80.20±0.23**
Ave		[59.98, 68.24]	70.01±1.82	68.82±2.27	70.93±2.39	69.95±1.80	70.29±1.92	71.75±1.89	**72.25±1.78**
WT			0	0	2	0	0	2	17

Table VII
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH BAYESIAN BASE CLASSIFIER (IN %).

Data	n	$[AC_L,\ AC_U]$	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	[90.61,93.17]	95.68±0.22	93.17±1.20	93.97±1.96	**95.82±0.00**	95.68±0.44	**95.82±0.01**	**95.82±0.01**
wq	5	[42.44,47.49]	46.89±0.12	47.94±2.75	**50.27±0.32**	46.75±0.00	47.64±0.98	49.24±1.06	49.24±1.06
wq	2	[44.65,46.83]	47.75±0.06	46.84±0.38	48.29±0.07	48.00±0.01	47.74±0.13	50.04±0.12	**50.19±0.01**
kn	2	[57.59,57.81]	83.38±2.06	83.61±3.03	81.59±2.41	83.33±0.00	83.64±3.12	84.18±0.89	**84.40±0.48**
Wb	6	[89.73,91.76]	93.13±0.19	92.88±2.28	93.58±0.13	93.32±0.00	93.41±0.43	94.06±0.12	**94.27±0.87**
Wb	3	[91.93,92.41]	93.27±0.50	92.88±1.75	92.97±1.35	93.67±0.00	93.23±0.69	93.06±0.62	**93.78±0.83**
rwq	5	[45.55,50.03]	56.14±0.19	54.83±1.72	56.97±0.84	56.22±0.00	56.58±0.04	57.09±0.44	**57.52±0.40**
rwq	2	[51.32,53.78]	56.41±0.19	56.29±1.07	55.54±0.38	56.85±0.00	56.33±0.99	58.19±0.44	**58.27±0.59**
pi	4	[65.29,74.97]	72.08±0.27	71.93±2.20	76.07±0.83	72.79±0.00	72.89±1.35	**76.17±0.75**	**76.17±0.75**
pi	2	[68.31,75.57]	75.39±0.00	75.29±1.93	**76.82±0.74**	75.39±0.00	75.68±1.30	76.02±0.92	76.30±0.75
tac	2	[43.27,45.04]	49.43±1.02	49.44±3.13	50.45±0.94	50.32±0.00	50.10±2.52	51.99±0.47	**53.42±1.37**
sat	6	[75.20,78.68]	80.39±0.11	79.32±0.75	80.46±0.09	80.31±0.00	80.35±0.03	80.42±0.23	**80.54±0.16**
sat	3	[78.36,79.43]	80.06±0.00	79.69±0.53	79.66±0.59	80.00±0.00	80.11±0.04	**80.20±0.26**	**80.20±0.26**
ma	5	[65.04,76.03]	73.25±0.10	72.06±0.90	**76.67±0.25**	73.43±0.00	73.39±0.16	76.33±0.07	76.43±1.08
ma	2	[70.94,74.39]	72.96±0.00	72.66±0.98	73.26±0.62	72.96±0.00	72.81±0.67	74.09±0.25	**74.37±0.72**
ve	6	[35.76,46.10]	46.22±0.67	43.68±3.59	50.06±3.93	45.51±0.00	45.86±0.67	53.96±0.08	**55.14±0.59**
ve	9	[34.55,44.09]	45.19±0.45	42.32±1.39	52.72±1.34	45.04±0.00	45.04±0.66	55.91±2.00	**56.50±1.45**
pb	5	[85.08,92.34]	90.91±0.58	90.20±0.74	91.81±0.43	93.01±0.01	91.22±0.01	93.42±1.09	**93.59±0.93**
te	8	[61.11,75.90]	76.14±0.30	76.01±1.94	78.57±0.55	77.45±0.01	77.46±0.01	79.78±0.72	**79.79±0.73**
te	4	[69.25,78.10]	77.01±0.78	73.94±1.56	77.36±0.72	77.45±0.00	77.05±0.14	77.65±0.23	**77.72±0.14**
Ave		[63.30, 68.70]	70.58±0.39	69.75±1.69	71.85±0.92	70.88±0.00	70.81±0.72	72.88±0.54	**73.18±0.66**
WT			0	0	3	1	0	3	17

Table VIII
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH HYBRID BASE CLASSIFIER (IN %).

Data	n	$[AC_L,\ AC_U]$	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	[89.76, 92.93]	93.68±2.07	93.77±2.02	93.12±2.45	93.68±1.82	93.59±2.14	93.86±1.81	**94.05±1.99**
wq	5	[44.33, 50.50]	46.41±0.23	47.14±2.28	**51.52±0.16**	46.78±0.13	47.12±0.38	50.55±1.79	50.70±2.01
wq	2	[43.17,48.69]	47.44±1.23	49.07±1.14	48.96±0.69	48.08±0.92	47.95±0.94	49.66±0.42	**49.86±0.66**
kn	2	[55.04,60.95]	78.56±5.18	82.27±4.91	83.18±3.89	78.23±5.90	84.16±6.47	84.15±4.95	**85.57±5.58**
Wb	6	[87.26, 92.88]	93.55±0.25	92.80±0.50	93.94±0.37	94.16±1.61	**94.55±0.74**	94.11±0.87	94.46±0.37
Wb	3	[89.28,92.70]	92.53±1.12	89.71±2.61	**93.23±0.62**	89.36±2.12	**93.23±0.62**	**93.23±0.62**	**93.23±0.62**
rwq	5	[39.65, 55.07]	56.66±0.18	56.07±0.40	57.54±1.94	56.54±0.00	56.97±0.62	57.63±1.10	**57.72±1.50**
rwq	2	[49.62,55.97]	55.88±2.17	57.26±0.75	57.22±0.44	56.50±0.84	56.38±1.46	57.66±0.88	**58.57±1.02**
pi	4	[65.17,75.33]	69.21±1.75	66.80±0.74	74.93±1.01	74.97±3.27	69.66±2.39	72.14±2.95	**75.13±1.29**
pi	2	[66.03, 73.26]	71.33±5.08	71.33±5.08	73.57±2.84	71.33±5.08	71.42±4.99	**74.41±3.06**	**74.41±3.06**
tac	2	[44.64,46.76]	50.33±3.14	47.69±3.57	51.52±1.29	49.67±1.93	50.93±3.43	52.32±1.62	**53.25±2.59**
sat	6	[58.90, 71.99]	75.56±0.27	69.31±1.69	67.57±2.75	75.18±0.24	75.63±0.08	75.45±0.00	**75.94±0.07**
sat	3	[71.58,74.05]	**76.04±0.24**	75.80±0.15	73.90±1.75	75.82±0.24	75.82±0.21	75.74±0.20	75.90±0.07
ma	5	[65.10, 73.93]	72.14±0.46	72.75±1.06	**75.34±1.20**	72.31±0.43	72.27±0.35	72.32±1.14	74.61±1.97
ma	2	[71.29, 74.84]	71.29±1.44	71.29±1.44	74.96±0.57	73.11±0.01	73.62±0.02	**75.18±0.70**	**75.18±0.70**
ve	6	[38.48, 53.13]	52.90±0.92	52.96±0.33	56.62±1.50	57.75±1.25	54.91±0.92	57.92±1.00	**58.70±1.60**
ve	9	[33.69,48.17]	53.49±4.26	44.68±2.60	56.15±1.00	46.22±5.68	53.15±4.10	60.64±2.17	**61.29±2.42**
pb	5	[88.53, 92.19]	89.81±0.68	89.37±1.03	90.17±0.85	90.82±1.56	90.02±0.75	91.34±1.27	**91.38±1.28**
te	8	[61.42, 78.48]	77.32±2.64	72.58±4.73	77.32±1.99	78.86±2.20	78.41±2.28	82.65±0.42	**83.55±0.22**
te	4	[68.25, 92.75]	73.22±0.10	73.90±0.48	75.45±1.98	75.57±0.22	79.12±2.28	82.23±1.48	**83.25±1.93**
Ave		[61.56, 70.23]	69.87±1.67	68.83±1.88	71.31±1.46	70.25±1.77	70.95±1.76	72.66±1.42	**73.34±1.55**
WT			1	0	3	0	2	3	16

Table IX
BENEFIT VALUE OF DIFFERENT COMBINATION METHODS WITH HYBRID BASE CLASSIFIER (IN %).

Data	n	AF	WAF	AFLW	DS	WDS	OWDS	OWDS-PW
new	2	93.88±2.11	94.11±2.09	93.07±2.29	93.88±1.83	93.88±2.01	94.07±1.80	**94.33±1.49**
wq	5	50.26±0.05	50.43±1.29	**52.55±0.24**	48.34±0.45	50.98±0.11	51.71±1.45	51.75±1.50
wq	2	49.83±1.38	51.08±0.06	50.99±0.05	50.05±0.69	50.00±0.59	51.73±0.67	**52.05±0.08**
kn	2	80.97±3.92	83.06±3.36	84.66±3.58	80.76±3.93	86.78±4.30	86.90±3.08	**87.95±4.67**
Wb	6	94.34±0.05	93.03±0.43	94.40±0.27	92.16±2.03	**94.71±0.51**	94.11±0.87	94.46±0.37
Wb	3	92.53±0.84	91.91±0.68	**93.44±0.98**	89.33±2.06	93.20±0.07	93.33±0.73	93.33±0.73
rwq	5	57.84±0.52	56.91±0.01	58.26±1.70	57.42±0.47	57.89±0.51	58.33±0.57	**58.70±1.06**
rwq	2	57.91±1.68	57.56±0.88	58.01±0.60	57.91±0.59	57.92±1.54	58.42±0.34	**58.81±0.53**
pi	4	70.82±3.30	69.05±2.29	75.33±1.83	69.74±2.26	70.86±3.49	72.52±2.41	**75.59±1.94**
pi	2	72.41±3.82	71.75±4.64	74.68±2.86	71.33±5.08	71.79±4.56	75.13±2.91	**75.15±2.78**
tac	2	53.68±3.47	50.38±1.95	54.93±0.89	50.98±2.45	54.61±2.57	55.73±1.59	**55.87±1.31**
sat	6	76.33±0.13	70.12±2.17	68.96±1.94	75.82±0.08	76.39±0.13	76.11±0.23	**76.62±0.10**
sat	3	76.74±0.11	76.50±0.02	74.75±1.96	76.70±0.12	76.72±0.11	76.73±0.03	**76.85±0.08**
ma	5	74.62±2.11	74.20±0.92	76.09±0.63	72.31±0.43	72.88±0.06	73.10±2.03	**76.26±2.67**
ma	2	73.98±1.21	73.62±1.12	75.60±0.39	73.11±0.01	74.46±0.58	75.78±0.51	**75.81±0.46**
ve	6	56.72±1.44	56.57±1.52	**60.61±1.40**	53.56±0.20	58.52±0.20	59.03±1.49	60.21±1.49
ve	9	58.11±1.98	49.38±1.93	59.21±1.02	48.23±6.04	57.39±1.51	60.71±1.40	**61.94±2.28**
pb	5	90.27±0.79	89.56±1.09	90.52±0.84	90.83±1.61	90.13±0.76	91.49±1.23	**91.54±1.26**
te	8	78.34±1.94	73.39±4.92	77.95±2.24	79.13±2.18	79.04±1.96	83.05±0.12	**83.92±0.47**
te	4	76.02±0.02	78.11±0.23	76.55±2.05	76.01±0.29	80.97±1.67	83.28±1.78	**84.50±2.37**
Ave		71.78±1.54	70.54±1.58	72.53±1.39	70.38±1.64	72.46±1.36	73.56±1.26	**74.28±1.38**
WT			0	3	0	1	0	16